REPORT ON SARC 41

JUNE 6–8, 2005

WOODS HOLE, MASSACHUSETTS

Prepared by

Patrick Cordue
Fisheries Consultant
New Zealand

for

University of Miami
Independent System for Peer Review

24 June 2005

EXECUTIVE SUMMARY

The 41st Stock Assessment Review Committee meeting was held June 6-8, 2005 at the Northeast Fisheries Science Center, Woods Hole, MA. Stock assessments of summer flounder, golden tilefish, and bluefish were reviewed by a Panel appointed by the Center for Independent Experts. This report details my individual findings on the quality of the stock assessments with regard to their adequacy to support management advice.

The three assessments were reviewed relative to each working group's terms of reference. The terms of reference for summer flounder included an assessment update and a re-evaluation of biological reference points (each using prescribed methods). For tilefish and bluefish, benchmark assessments were presented. Tilefish was last assessed in 1999. The bluefish assessment was unusual in that the previous assessment had been rejected only one year ago at SARC 39.

The update of the summer flounder assessment and the revision of biological reference points were done according to the working group's terms of reference and are adequate as a basis for providing scientific advice to management. However, the interpretation of the retrospective pattern seen in the current assessment is incorrect and wording in the assessment documents should be changed to reflect a correct interpretation.

The bluefish assessment was technically deficient and inadequately documented. The nature of the bluefish data required that a careful statistical modeling exercise be undertaken. This was outside the expertise of the Bluefish Technical Committee. They correctly adopted the use of a statistical model (ASAP) but because of inexperience with such models were unable to produce a statistical assessment of sufficient quality on which to base management advice. There are several deficiencies in the bluefish assessment: using age-length data from adjacent years to fill in a catch-at-age matrix; using state surveys (with limited spatial coverage) as population indices; using catch-at-age data (already input to the model) to produce recreational CPUE-at-age; inappropriate determination of the relative weights of indices; an absence of any measure of uncertainty for the ASAP model results; an absence of ASAP sensitivity runs (apparently done but not documented); no objective method of estimating coefficients of variation and effective sample sizes; no documentation of assumed coefficients of variation and effective sample sizes; and an absence of meaningful diagnostics for ASAP model fits.

The tilefish assessment and the revision of biological reference points were done according to the working group's terms of reference and are adequate as a basis for providing scientific advice to management. A wide range of sensitivities were tested and the assessment results appear quite robust.

There is a need to move to the use of data-appropriate (statistical) stock assessment models in the short to medium term for many species. Statistical modeling requires a major shift in philosophy and technique from that applied in virtual population analysis. Assessment scientists more familiar with the latter than the former need more exposure to statistical modeling ideas and techniques and need to gain experience in applying the methods *before* using them in benchmark stock assessments. I suggest, if it is not already in place, that a staff training plan be developed, including workshops on: data preparation (including estimation of CVs and effective sample sizes); modeling tools (e.g., stock synthesis II); and stock assessment "best practice" (e.g., relative weighting of time series, model diagnostics, capturing uncertainty).

BACKGROUND

The 41st Stock Assessment Review Committee (SARC) meeting was held June 6-8, 2005 at the Northeast Fisheries Science Center, Woods Hole, MA. Three stock assessments (summer flounder, golden tilefish, and bluefish) were reviewed by a Panel appointed by the Center for Independent Experts (CIE). The SARC was run under the "new model" which was first used for SARC 39. There are four main differences between the new and old processes. First, the Panel reviewing the assessments was previously quite large (12-18 people) and included only two CIE appointees. The Panel is now entirely CIE, with a Chair and three Panelists (see Appendix 4). Second, a consensus opinion on the quality of the assessments is no longer sought. The panelists report individually and the Chair summarizes their *individual* opinions. Third, the SARC used to provide management advice, but now the assessment documents are primarily aimed at summarizing the "status of the stocks". Finally, the reports provided by the CIE Panelists are included in the SARC Assessment Report and are made public. Previously, the CIE reports were not widely distributed.

The three assessments were reviewed relative to each working group's terms of reference (TOR, see Appendix 3). The Southern Demersal Working Group (SDWG) assessed summer flounder and golden tilefish. The TOR for summer flounder were quite narrow but included an assessment update (using the methods and formulation of the previous assessment) and a re-evaluation of biological reference points (methods specified). The tilefish TOR were somewhat broader. A "benchmark" assessment was required. Data were to be evaluated and an assessment produced using whatever methods the SDWG deemed appropriate. The bluefish assessment was conducted by the Atlantic States Marine Fisheries Commission Bluefish Technical Committee (BTC). This was also a benchmark assessment but it was unusual in that the previous assessment had been rejected only one year ago at SARC 39.

REVIEW ACTIVITIES

Meeting Preparation

Documents were available electronically from a website and were provided in hardcopy well in advance of the meeting. Prior to the meeting I read the assessment documents and consulted the supplied background material (Appendix 1). I also attended an informal pre-meeting briefing where the SAW Chair provided some very useful background information on the new SAW/SARC process.

Meeting Attendance

A narrative of the meeting is given below. These details are aimed at providing the reader with an understanding of what took place at the meeting – at least from my perspective – and a context for my subsequent findings with regard to the assessments.

6 June

The meeting was convened at 1pm and began with an introduction from the SAW Chair followed by a round of introductions. The SAW Chair then described the new SAW/SARC process (as of

SARC 39). The agenda was briefly reviewed. The first presentation was on Summer flounder by Mark Terricio.

The TOR for summer flounder were listed and each was covered in turn. There were few substantive questions. Age-length keys for use with recreational length frequencies had been constructed from commercial and survey data. Given the use of different methods and the different depth ranges covered, this was of concern to me. However, people more familiar with the fish and the fisheries thought it unlikely to be a problem. The precision of the catch-at-age estimates was an issue. No estimates of precision had been made. Of course, in the VPA they are assumed to be exact. The use of numerous state surveys as tuning indices was raised. Given their limited spatial coverage they cannot be expected to provide unbiased population indices. However, according to the TOR, a prescribed formulation of the VPA (which included the state surveys) had to be used.

I had an issue with regard to the interpretation of the results from the retrospective analysis. During the presentation, Terricio was careful to refer to a "retrospective bias" (rather than just "bias"). However, in the Assessment Summary document there were several sentences that implied that the current estimators of biomass and fishing mortality were biased, as evidenced by a retrospective analysis. A Panelist noted that in a herring assessment, where a retrospective pattern was present, the current "bias" had been corrected for by using the difference between converged and retrospective estimates. I gave a short presentation entitled "Retrospective analysis, a note on correct interpretation" (Appendix 2). The issue was discussed by the meeting: what does a retrospective pattern actually indicate, if anything? I indicated that retrospective analysis, as currently practiced, only dealt with a single realization of a time series of random variables. As such, the results cannot reveal anything about the bias or variance of the estimators of current biomass or current fishing mortality (see Appendix 2).

The presentation on the recalculation of biological reference points produced almost no comments from the SARC Panel. The results, which had been considered plausible by the SDWG, were all very similar. The SDWG favored the updated estimates provided by the previously used method. The Panel did not object. A Panelist asked who set the TOR for a WG, and in particular for Summer flounder. It seemed odd that we were reviewing an assessment update, rather than a "benchmark" assessment. Apparently, the main purpose of the review had been to consider the new biological reference points. The meeting closed for the day at 5pm

7 June

The morning session was devoted to bluefish. An excellent presentation was given by Jessica Coakley, supported by Gary Shepherd. It was refreshing to be given useful information on the biology of the species and to be shown maps illustrating survey coverage and results, relative to the stock's distribution. Interesting features of bluefish include the appearance of spring and summer cohorts in the length frequencies and the lack of medium sized fish, especially in the commercial fishery. Migration patterns were discussed, but a graphic illustrating the main hypotheses and their relevance to abundance surveys would have been useful.

There was some discussion of the previous assessment rejected by SARC 39. There were two main problems: an error made in the calculation of the recreational CPUE indices (dead fish used rather than total encounters); and the sensitivity of the model estimates to starting values in the minimization. After completion of the presentation, the Panelists gave their assessment of the adequacy of the stock assessment for management advice. John Wheeler and Olav Godø appeared to have relatively minor concerns about the available data and the absence of "older fish"

respectively. This put me in the position of having to point out the many flaws in the assessment methodology and the deficiencies of the assessment documentation. I posed the question as to whether I should adopt different standards in different parts of the world, and make allowances for the steep learning curve when moving from VPAs to statistical assessment methods. I was assured that I must maintain a consistent standard – whatever it may be – when judging the scientific merits of assessments. That said, I detailed my concerns: using age-length data from adjacent years to fill in a catch-at-age matrix; using state surveys (with limited spatial coverage) as population indices; using catch-at-age data (already input to the model) to turn the recreational CPUE time series into CPUE-at-age; determining relative weights of indices in a VPA and then using these in the ASAP model; an absence of any measure of uncertainty for the ASAP run; absence of sensitivity runs (apparently done but not documented); no objective method of estimating coefficients of variation (CVs) and effective sample sizes; no documentation of input CVs and effective sample sizes assumed; and the absence of meaningful diagnostics for the ASAP model (e.g., standardized residuals; analysis of residual patterns).

The afternoon session was devoted to tilefish, presented by Paul Nitschke (main presentation) and Jon Brodziak (alternative model). This is entirely a commercial fishery that recently has only involved about 10 vessels using longlines (the fish live in burrows). The main assessment runs used ASPIC and there were many sensitivity runs done, including using alternative models (lagged recruitment survival growth model; and a catch-length forward projection model). I briefly introduced the notion of a fully integrated model using the total catch, CPUE time series, and the length frequency data – but discussion was difficult because of a general lack of understanding as to how such models could deal with the multiple data sources and apparent data conflicts.

The Panel was unanimously positive with regards to the tilefish stock assessment. Also, with so few vessels in the fishery, we agreed that there is a good opportunity for cooperative research with comprehensive data sampling being entirely feasible. We had an early finish at 4 pm.

8 June

The morning began with a "closed session" where the Panel discussed with the SAW Chair which species should be revisited. The bluefish team wished to talk to us in terms of offering a "rebuttal" to the criticisms of their assessment. Mark Terricio wished to clarify with us some possible changes to the summer flounder summary report (I had suggested some changes – mainly editorial in nature). We asked to see Paul Nitschke, to obtain clarification on a few minor issues. Also, there was some discussion on the nature of our recommendations. We concluded that they should be feasible, prioritized, and categorized as short, medium, or long term.

The session for the bluefish discussion was well attended. Shepherd presented some arguments in support of the way the assessment had been done. The use of age data from adjacent years to fill out age-length keys was explained on the basis that most fish in the catch were quite young. The proportion at age for a given length was therefore well determined for the smaller fish. Unfortunately, it was not only age data for small fish that was borrowed from other years. There was robust discussion about other aspects of the assessment where I had been critical. I was asked if my main concern was the documentation of the assessment or the methods used. I indicated that the method was technically flawed *and* the documentation was inadequate.

The summer flounder assessment summary was considered next. Terricio had three issues for us to consider. The first was a relatively minor question about the correct terminology for the bootstrap confidence intervals that had been constructed. Apparently, there had been advice from

previous SARCs, with suggestions of "probability intervals", "confidence intervals", "bootstrap confidence intervals" and assertions that they were invalid because they had not been "bias corrected". This SARC (the Chair and myself) were adamant that statements must not refer to "probability" or "chance"; the intervals were "confidence intervals". This change was adopted in summer flounder (and also in tilefish).

The second issue concerned several statements in the summary report that referred to tendencies to "underestimate recent fishing mortality rates", based on the observed retrospective pattern. Indeed, there is a recommendation in the report that managers should consider adopting a lower TAL for 2006 than specified by the assessment, given the retrospective pattern. It was clear that any changes to these statements were beyond editorial and so could not be made (it was for me to comment on them in my report). The third issue concerned whether the new reference points should be used (instead of the previous reference points) in comparisons with the estimated biomass and fishing mortality. The Panel agreed that the new reference points should be used.

There was a brief discussion on tilefish with Nitschke. I expressed two concerns about issues that had not been fully discussed in the previous session. First, there was the issue of the current stock hypothesis. As it appears to be based almost entirely on the separation of commercial fishing grounds, I suggested that the issue might need to be revisited. Second, as the assessment results depended so strongly on the CPUE time series, I suggested that more work might have been done exploring possible effort trends that may have compromised the validity of the time series as an abundance index (e.g., an increase in the number of hooks per day). There was also some discussion on the length frequency data from the fishery. Cohorts could be "tracked" from year to year in the data, and the 1999 cohort appeared to be strong. However, there was no indication of the cohorts since 1999 in the data. A fishing industry representative at the meeting indicated that some vessels had shifted to larger hooks in recent years to avoid catching the smaller fish.

Conduct of the Meeting

The meeting was held in a constructive and amicable atmosphere. Presentations were professional and discussions informative. The Chair adopted a relaxed style which, given the participants, allowed the meeting to progress efficiently.

Strengths and Weaknesses in Assessment Methods and Advice

Summer flounder

The SDWG completed the assessment update as per their TOR. They used a VPA with the formulation used in the previous assessment. Bootstrap confidence intervals were calculated for derived outputs of interest. I have some minor concerns with regard to the assessment. The catch-at-age data for summer flounder appear reasonably good (in terms of sample sizes and coverage) but nevertheless probably fall far short of providing the very accurate estimates that are required by a VPA. I also note that the level of precision has not been estimated (see Bull and Dunn 2002 for a suitable method). The age-length keys for the recreational catch are obtained from commercial and survey data. They may provide a reasonable approximation to what would be obtained from the recreational catch, but there are no data to test this. The use of the individual state surveys is obviously suspect given their limited spatial coverage. Finally, the bootstrap

confidence intervals will underestimate uncertainty because they do not account for error in the catch-at-age matrix.

The estimation of biological reference points was done according to the TOR. The SDWG tried a range of methods that produced a set of plausible estimates that showed little variation (there were other estimates that were not plausible – caused by clear failure of the method used). I have not considered the technical merits or deficiencies of the different methods in detail as the methods were prescribed in the TOR. I suggest that some future consideration be given to risk-based definitions of reference points (e.g., Francis 1992).

From my perspective, the main issue for summer flounder is the misinterpretation of the results of the retrospective analysis. The wording in the assessment documents imply that the current estimators of fishing mortality and biomass are biased. They may well be, but a retrospective analysis cannot demonstrate a statistical bias nor can it be used to accurately estimate its direction or magnitude. This is because a retrospective analysis only considers a single realization of a set of observations. A particular realization of a catch-at-age matrix and tuning indices may produce a "retrospective bias" of some direction and magnitude. However, another realization may give rise to a completely different pattern with a different "bias" – or no "bias" at all (see Appendix 2).

It may be the case that retrospective patterns are more common if there is model "mis-specification" in a VPA. However, if there are errors in the model (e.g., catch-at-age systematically under-estimated, or assumed natural mortality too low) then it would be imprudent to assume that the converged estimates are any more reliable than the non-converged estimates. In either case, model mis-specification or not, the difference between converged estimates and non-converged estimates (for previous years) is not generally indicative of the magnitude or direction of an error in the current estimates. (I have no compelling literature to cite in support of my assertion. I am considering drafting a short communication, with VPA examples, that will illustrate the point. In the interim, consider Appendix 2 and see Cordue 2004, Appendix D.)

Bluefish

The BTC evaluated several models before choosing ASAP as their primary assessment model (B 1, Appendix 1). The data were formulated for use with an ADAPT VPA. This included constructing a catch-at-age matrix for a continuous number of years despite inadequate samples of otoliths from 1997 to 2004 (see B-1, Table 8). The re-weighting of indices was done within the VPA model. However, the BTC realized that the VPA assumption of exact catch-at-age was not appropriate and that ADAPT's treatment of fishery selectivity was inadequate for bluefish. Therefore, they moved to ASAP to allow errors in the catch-at-age and more flexibility in the fishery selectivity pattern. The move to a more data appropriate model was laudable but unfortunately the data were not reformulated and statistical error structures were barely considered. The use of statistical models requires a major shift in philosophy from that associated with VPAs. It was unfortunate that the BTC had little or no experience with statistical modeling.

The first major deficiency in the (ASAP) assessment was borrowing data from adjacent years to construct age-length keys so that a continuous catch-at-age matrix could be estimated. It may be that ASAP requires a catch-at-age matrix without missing years. If this is the case, it is a deficiency in ASAP. In statistical modeling the idea is to provide the model with genuine data, not to construct data to fill gaps – the gaps are genuine and must be allowed to contribute to the uncertainty of the assessment.

The use of state surveys as population indices is another deficiency in the assessment. With such limited spatial coverage their use as population indices cannot be justified. Clearly, there is a need (for many species) to develop a method of combining state indices so that the data can be used in stock assessments for coast-wide stocks.

Appropriate use of data is fundamental to statistical modeling. Using "data twice" is best avoided, if at all possible, because it will introduce correlations between time series that are not accounted for in the likelihood equations. Unfortunately, the bluefish recreational catch-at-age data were used twice: as a (major) component of the total catch-at-age matrix and to split the recreational CPUE indices into CPUE-at-age. Such an approach emphasizes the recreational catch-at-age, giving it more "weight", relative to other indices, than it should have. However, it is not clear that any indices had appropriate "weight" in the fitting procedure (assumed CVs and effective sample sizes are not given in the assessment documents). No consideration was given to varying the weight of indices across years within time series (despite variations in sample sizes). Also, the relative weights of time series in the ASAP model were primarily determined using ADAPT; any re-weighting of time series clearly should have been done in ASAP.

The lack of meaningful diagnostics in the bluefish documentation is also of concern. Claims of a "good fit" to indices cannot be substantiated without the production of standardized residuals that account for the weight placed on the indices. The tabulation of likelihood values and residual sum of squares is not useful for a single run (B-1 Table 22). Comparison of likelihood values across runs can be useful, but no ASAP sensitivity runs are presented in the assessment documents. Finally, no estimates of uncertainty were presented for the ASAP base model. Bootstrap confidence intervals were calculated for the ADAPT run but these are obviously not transferable to the ASAP run.

Tilefish

The SDWG considered four different models in a thorough stock assessment of tilefish. For their base assessment model (ASPIC) they did extensive sensitivity runs. The SDWG rightly expressed concern that the assessment had to rely on a commercial CPUE time series that may not be tracking population abundance. However, the sensitivity runs adequately explored the robustness of the assessment results to this assumption. Particular care was taken to explore the sensitivity of results to the large increase in the CPUE indices in recent years.

The use of a range of models in this assessment was meritorious, but no single model was able to use all of the available data. There appear to be signals with regard to cohort strength in the length frequency data. These data were only used in the catch-length model, which was not able to use an abundance time series, and detected an apparent conflict between the catch history and the commercial length frequencies. A fully integrated model, which can use all of the available data, should be tried for tilefish. There is no need to develop new tools in this regard as suitable packages already exist (e.g., SS2, Methot, drafts a & b; CASAL, Bull et al. 2003). It is by using an integrated model that apparent conflicts in data can be explored and possibly resolved. One must be aware that data are never contradictory by themselves. Data sets may appear to be contradictory but such appearances are always contingent on certain model assumptions – alter the assumptions and apparent conflicts may disappear.

SUMMARY OF FINDINGS

Summer flounder

The update of the summer flounder assessment and the revision of biological reference points were done according to the TOR and are adequate as a basis for providing scientific advice to management. However, the interpretation of the retrospective pattern seen in the current assessment is incorrect and wording in the assessment documents should be changed to reflect a correct interpretation.

Bluefish

The bluefish assessment is not adequate as a basis for providing scientific advice to management. There are several deficiencies in the assessment:

- using age-length data from adjacent years to fill in a catch-at-age matrix;
- using state surveys (with limited spatial coverage) as population indices;
- using catch-at-age data (already input to the model) to produce recreational CPUE-at-age;
- determining relative weights of indices in the VPA and then using these in the ASAP model;
- an absence of any measure of uncertainty for the ASAP model results;
- absence of ASAP sensitivity runs (apparently done but not documented);
- no objective method of estimating CVs and effective sample sizes;
- no documentation of assumed CVs and effective sample sizes;
- absence of meaningful diagnostics for ASAP model fits (e.g., standardized residuals; analysis of residual patterns).

The data available for a bluefish assessment are adequate for use in a statistical assessment model. However, the data must be used appropriately, with due consideration of appropriate CVs, effective sample sizes, and error structures. Good practice needs to be followed with appropriate sensitivity runs, model diagnostics, and estimates of uncertainty.

Tilefish

The tilefish assessment and the revision of biological reference points were done according to the TOR and are adequate as a basis for providing scientific advice to management. A wide range of sensitivities were tested and the assessment results appear quite robust.

CONCLUSIONS AND RECOMMENDATIONS

The summer flounder and tilefish assessments satisfied their TOR and the assessments provide an adequate basis for the provision of management advice. The bluefish assessment was technically deficient and inadequately documented. This was not the fault of the Bluefish Technical Committee. The nature of the bluefish data required that a careful statistical modeling exercise be undertaken. This was outside the expertise of the Bluefish Technical Committee. They correctly

adopted the use of a statistical model (rather than a VPA) but, due to inexperience, were unable to produce a statistical assessment of sufficient quality on which to base management advice.

I have three generic recommendations that are applicable to many species including those considered by the SARC. I see each of the recommendations as co-requisites for the necessary shift to the use of data-appropriate (statistical) stock assessment models in the short to medium term.

1. Statistical modeling requires a major shift in philosophy and technique from that applied in VPA assessments. Assessment scientists more familiar with the latter rather than the former need more exposure to statistical modeling ideas and techniques and need to gain experience in applying the methods *before* using them in benchmark stock assessments. I suggest:

- a staff training plan be developed, including workshops on:
    - data preparation (including estimation of CVs and effective sample sizes);
    - modeling tools (e.g., SS2);
    - stock assessment "best practice" (e.g., relative weighting of time series, model diagnostics, capturing uncertainty).

2. Statistical modeling does not require estimates of catch-at-age from all modes of fishing for every year. However, it does require that what data there are do satisfy the statistical assumptions of the model. This invariably means that catch samples must be representative and appropriately stratified. The level of sampling is not nearly as important as the need for representative sampling. I suggest that:

- catch sampling schemes be reviewed for a wide range of species;
- consideration be given to how state sampling can be coordinated (across states and species);
- species need to be prioritized, not all species can be sampled every year;
- careful consideration be given to sampling protocols with regard to representative/random sampling;
- quality is more important than quantity.

3. State surveys could provide a valuable source of data for many species. However, the surveys need to be used appropriately taking account of the spatial and temporal distribution of the species relative to the coverage of the surveys. Four steps are needed:

- create a meta database for the surveys (what data is available and who has it);
- create a database to hold all of the state survey data (i.e., raw data, all in one place);
- develop methods for producing abundance indices for species by combining appropriate state surveys;
- apply the methods to produce abundance time series for all suitable species.

I also have some species specific recommendations detailed below.

Summer flounder

Short term goals:

- Reconsider the interpretation of the retrospective pattern and revise the assessment documents accordingly.
- Design an experiment to test the assumption that the current age-length keys are adequate for use with the recreational length frequencies.

Medium term goal:

- Develop a fully integrated stock assessment model that uses the existing data appropriately.

Bluefish

Short term goals:

- Formulate stock and migration hypotheses, detailing the basis for each hypothesis.
- To the extent possible, test the alternative hypotheses (e.g., using tagging data)
- Review ageing data:
    - comparison of scale and otolith ageing
    - accuracy of bluefish ageing (between and within reader comparisons).

Medium term goal:

- Develop a fully integrated stock assessment model that uses the existing data appropriately.

Tilefish

Short term goals:

- Formulate stock and migration hypotheses, detailing the basis for each hypothesis.
- Review CPUE data with a view to testing for possible changes in effective effort (e.g., an increase in the number of hooks per day).

Medium term goals:

- Develop cooperative research with the industry participants, including:
    - examining/capturing existing vessel/skipper logbook data
    - implementing a comprehensive scientific logbook scheme with biological sampling and fine-scale CPUE data
- Develop a fully integrated stock assessment model that uses the existing data appropriately.

REFERENCES

(see Appendix 1 for further references)

Bull, B.; Dunn, A. (2002). Catch-at-age: User manual v1.06.2002/09/12. NIWA Internal Report 114. 23 p. (Unpublished report held in NIWA library, Wellington.)

Bull, B.; Francis, R.I.C.C.; Dunn, A.; McKenzie, A.; Gilbert, D.J.; Smith, M.H. (2003). CASAL (C++ algorithmic stock assessment laboratory): CASAL User manual v2.01-2003/8/01. *NIWA Technical Report 124*. 223 p.

Cordue, P.L. (2004). An evaluation of the management utility of fishery performance indicators commonly used in New Zealand stock assessments. *New Zealand Fisheries Assessment Report 2004/23.* 72 p.

Francis, R.I.C.C. (1992). Recommendations concerning the calculation of maximum constant yield (MCY) and current annual yield (CAY). New Zealand Fisheries Assessment Research Document 92/8. 27 p. (Unpublished report held in NIWA library, Wellington.)

Methot, R.D. (draft a). User Manual for the Assessment Program Stock Synthesis 2 (SS2) Model Version 1.17. Draft document dated 4 April 2005.

Methot, R.D. (draft b). Technical Description of the Stock Synthesis II Assessment Program Version 1.17. Draft document dated March 2005.

APPENDIX 1: MATERIAL PROVIDED

General Documents

Revisions to the Northeast Regional Stock Assessment Workshop –'old' versus "new".

Terms of Reference for the 41st Northeast Stock Assessment Workshop, (approved March 18, 2005), SAW/SARC 41, June 6-10, 2005, NEFSC, Woods Hole, MA.

Summer Flounder

A-1 Summer flounder: Stock assessment update and biological reference point estimation by SAW Southern Demersal Working Group, Mark Terceiro, NMFS/NEFSC.

A-2 Summer Flounder Appendix A: Data Tables & Figures.

A-3 Summer Flounder Appendices B: 1) ADAPT VPA Output, 2) AGEPRO Projection Output.

A-4 SSC Committee Overfishing Definition (2001).

A-5 A: Summer Flounder, SAW/SARC-35 Report (2002), NEFSC Reference Document 02-14.

A-6 Stock Assessment of Summer Flounder for 2003 by Mark Terceiro, August 2003, NEFSC Reference Document 03-09.

A-7 SAW Southern Demersal Working Group 2004 Summer Flounder Assessment Summary, June 21, 2004.

A-8 Re-evaluation of biological reference points for New England groundfish by Working Group on Re-Evaluation of biological reference points for New England groundfish, March 2002, NEFSC Reference Document 02-04.

SARC/SAW-41 Two PowerPoint presentations by Mark Terceiro, June 6, 2005.

SARC/SAW-41 Summer Flounder Rapporteur Report by Kathy Sosebee.

Bluefish

B-1 B: Working paper for bluefish stock assessment 41st Northeast Stock Assessment Workshop working document for Stock Assessment Review Committee, June 6-10, 2005.

B-2 Corrections to Paper B1: Bluefish SAW-41 Working Group Stock Assessment Report (May 24, 2005).

B-3 SARC 41: Bluefish Assessment Summary (WG draft 5/20/05).

B-4 C: Bluefish SARC Report SAW/SARC-23 (1996).

B-5 Report on the 39th Northeast Regional Stock Assessment Workshop (SAW-39) Stock Assessment Review Committee (SARC) Meeting by Andrew I.L. Payne, SARC-39 Chair.

SARC/SAW-41 PowerPoint presentation by Jessica Coakley, June 7, 2005.

SARC/SAW-41 Bluefish Rapporteur Report by Gary Shepherd.

Tilefish

C-1 Assessment of golden tilefish, *Lopholatilus chamaeleonticeps*, in Middle Atlantic-Southern New England Region, SAW 41 SARC Working Paper C1, a report of the Southern Demersal Working Group, NMFS/NEFSC.

C-2 Golden Tilefish Summary Report, SARC 41.

C-3 Assessment of tilefish in the Middle Atlantic-Southern New England Region by Paul Nitschke, Gary Shepherd, and Mark Terceiro (1998) for S&S Committee Review.

C-4 G. Tilefish (Background SAW/SARC 16).

SARC/SAW-41 PowerPoint presentation by Paul Nitschke, June 7, 2005.

SARC/SAW-41 PowerPoint presentation by John Brodziak, June 7, 2005.

SARC/SAW-41 Tilefish Rapporteur Report by Laurel Col.

APPENDIX 2: A NOTE ON THE CORRECT INTERPRETATION OF RESULTS FROM RETROSPECTIVE ANALYSIS

The slides from the PowerPoint presentation that I presented to the meeting are given below. I use the convention of capital letters for random variables and lowercase letters for particular realizations. The definition of retrospective analysis is kept simple for illustrative purposes (no reference is made to what is being estimated).

Definitions

- Available data: $\{ x_1, x_2, \ldots, x_n \}$
- $y_j = \{ x_1, x_2, \ldots, x_j \}$
- G an estimator:
    - $g_j = G(y_j)$
- Retrospective analysis, consider estimates:
    - $g_n, g_{n-1}, g_{n-2}, \ldots$

Common interpretation

- Any trend in $g_n, g_{n-1}, g_{n-2}, \ldots$is bad indicating a problem with the estimator G_n
- Lack of trend is good, suggesting G_n is a good estimator
- VPA: consistent difference between non-converged estimates and converged estimates indicates a bias

However, …

- The standard interpretations all involve conclusions about the estimator G_n
- Properties of G_n, such as bias, can only be determined analytically (sometimes) or by simulation.
- They cannot be determined from a single realization of $\{X_1, \ldots, X_n \}$

Examples

- Toss a coin n times, estimate probability of success with $\sum X_i/n$
- Three different realizations, n = 6:
    - 1, 1, 1, 0, 0, 0 (1, 1, 1, 0.75, 0.6, 0.5)
    - 0, 0, 0, 1, 1, 1 (0, 0, 0, 0.25, 0.4, 0.5)
    - 1, 0, 1, 0, 1, 0 (1, 0.5, 0.67, 0.5, 0.6, 0.5)
- Common interpretation of retrospective analysis results gives a different conclusion about the estimator for each realization

Are VPAs different?

- If all assumptions of the model are satisfied then converged estimates are accurate estimates of the true values
- But the retrospective pattern could be a product of the particular realization of the tuning indices
- In any case, the difference between an estimate and the true value is an error not a bias (which is a property of an estimator)

What does the summer flounder retrospective pattern mean?

- Weakly suggestive that when current estimates are converged that they may show higher F and lower biomass
- If VPA model assumptions are satisfied, then it is weakly suggestive of an error in current estimates – there may be a bias, but simulations are needed to find out
- If VPA model assumptions are not satisfied then no valid conclusions on the nature of the errors can be made

APPENDIX 3: TERMS OF REFERENCE FOR THE 41ST NORTHEAST STOCK ASSESSMENT WORKSHOP

(approved: March 18, 2005)

SAW/SARC 41
June 6-10, 2005
NEFSC, Woods Hole, MA

Bluefish - ASMFC Technical Committee/Assessment Subcommittee

1. Evaluate adequacy, appropriateness and uncertainty of fishery-dependent and fishery-independent data used in the assessment.
2. Evaluate adequacy and appropriateness of models used to assess the species and to estimate population benchmarks.
3. Evaluate and either update or re-estimate biological reference points as appropriate.
4. Estimate and evaluate stock status (biomass) and fishery status (fishing mortality rate).
    a. Is the stock overfished; is overfishing occurring?
5. Develop recommendations for improving data collection and for future research.

Tilefish - SAW Southern Demersal Working Group

1. Characterize the commercial catch including landings and discards. Characterize recreational landings.
2. Estimate fishing mortality and total stock biomass for the current year and characterize the uncertainty of those estimates.
3. Evaluate and either update or re-estimate biological reference points as appropriate.
4. Where appropriate, estimate a constant TAC and/or TAL based on stock status for years following the terminal assessment year.
5. If projections are possible,
    a. provide seven year projections of stock status under various TAC strategies and
    b. evaluate current and projected stock status against existing rebuilding or recovery schedules, as appropriate.
6. Review, evaluate and report on the status of the research recommendations offered in the 1999 Science and Statistical committee reviewed assessment.

Summer Flounder - SAW Southern Demersal Working Group

1. Update the summer flounder assessment models (i.e. ADAPT VPA and AGEPRO projection) using the same configurations as those used in the 2004 SAW Southern Demersal Working Group (WG) assessment update.
2. Estimate biological reference points derived by yield and SSB per recruit analysis and by stock-recruitment modeling, following the procedures adopted by the 2002 Working Group on Re-Evaluation of Biological Reference Points for New England Groundfish.
3. Consider the recommendations of the MAFMC Science and Statistical Committee (SSC) 2001 peer review of the summer flounder Overfishing Definition in developing the analyses described in TOR 2. The major recommendations were to explore other proxies (besides

F_{max}) to F_{MSY}, to continue stock-recruitment model development as additional stock-recruit estimates become available, and to monitor and utilize new data on the population dynamics of summer flounder (e.g., age, growth, and maturity) as they become available.

4. Review, evaluate and report on the status of the SARC/Working Group research recommendations offered in previous SARC and WG reviewed assessments.

SARC issues

I raised two issues not directed related to any of the Working Group's TOR:

1. The correct interpretation of results from retrospective analysis.
2. The correct terminology for describing interval estimates obtained from bootstrapping.

APPENDIX 4: STATEMENT OF WORK

Statement of Work
Consulting Agreement between the University of Miami and Patrick Cordue

May 5th, 2005

GENERAL

The Northeast Regional Stock Assessment Review Committee meeting (SARC) is a formal, multiple-day meeting of stock assessment experts who serve as a peer-review panel for several tabled stock assessments. The SARC is the cornerstone of the Northeast Stock Assessment Workshop (SAW) process, which includes peer assessment development (SAW Working Groups or ASMFC technical committees), assessment peer review, public presentations, and document publication.

The Center for Independent Experts (CIE) shall provide a panel chair and three panelists for the 41st Stock Assessment Review Committee panel. The panel will convene at the Woods Hole Laboratory of the Northeast Fisheries Science Center in Woods Hole, Massachusetts, the week of 6 June 2005 (June 6-10) to review assessments for bluefish (*Pomatomus saltatrix*), tilefish (*Lopholatilus chamaeleonticeps*) and summer flounder (*Paralichthys dentatus*).

Specific Activities and Responsibilities

The CIE's deliverables shall be provided according to the schedule of milestones in the table below. The final reports from the CIE will provide key information for a presentation to be made by NOAA Fisheries at meetings of the New England and Mid-Atlantic Fishery Management Councils in August and September 2006. The chair's duties shall occupy a maximum of 19 days (i.e., several days prior to the meeting for document review; the SARC meeting in Woods Hole; and several days following the meeting to review the individual panelist's Review Reports and produce the Summary Report). This report shall be a summary of the individual Review Reports, accurately and fairly representing all viewpoints. There shall be no attempt by the Chair to develop a consensus report.

Each panelist's duties shall occupy a maximum of 14 workdays (i.e., a few days prior to the meeting for document review; the SARC meeting; and a few days following the meeting to prepare a Review Report). The SARC Review Reports will be provided to the SARC Chair, who will produce the Summary Report based on the individual Review Reports.

Roles and responsibilities:

(1) (Chair and Panelists) Prior to the meeting: review the reports produced by the Working Groups.
(2) (Panelists) During the meeting: participate, as a peer, in panel discussions on assessment validity, results, recommendations, and conclusions especially with respect to the adequacy of the assessments reviewed in serving as a basis for providing scientific advice to management.

(3) (Panelists) After the meeting: prepare individual Review Reports, each of which provides an executive summary, a review of activities and, for each stock assessment reviewed, a summary of findings and recommendations that emerge from the findings, all in the context of responsiveness to the Terms of Reference for each assessment. Advice on additional questions that are directly related to the assessments and are raised during the meeting should be included in the report text. These additional topics/issues should be listed along with the original Terms of Reference in a separate appendix attached to the report. See Annex 1 for further details on report contents and milestone table below for details on schedule. No later than June 24, 2005, these reports shall be submitted to the CIE for review[1] and to the Chair for summarization. The CIE reports shall be addressed to "University of Miami Independent System for Peer Review," and sent to Dr. David Sampson, via e-mail to David.Sampson@oregonstate.edu and to Mr. Manoj Shivlani via e-mail to mshivlani@rsmas.miami.edu.

NEFSC staff and the SAW Chairman will be responsible for the production of the final SARC report, which will include the Chair's Summary Report and the individual panelist's Review Reports. Staff and the SAW Chairman will also be responsible for production and publication of the collective Working Group papers, which will serve as a SAW Assessment Report.

NEFSC Contact person and SAW41 Chairman:
Dr. James R. Weinberg, NEFSC, Woods Hole, MA. 508-495-2352,
James.Weinberg@noaa.gov

[1] All reports will undergo an internal CIE review before they are considered final.

ANNEX 1: Contents of Panelist Report

1. The report shall be prefaced with an executive summary of findings and/or recommendations.

2. The main body of the report shall consist of a background, description of review activities, summary of findings, conclusions/recommendations, and references.

3. The report shall also include as separate appendices the bibliography of all materials provided during SAW 41 and any papers cited in the Panelist's Report, along with a copy of the statement of work.

4. The report shall also include as a separate appendix the Terms of Reference used for SAW 41, including any changes to the Terms of Reference or specific topics/issues directly related to the assessments and requiring Panelist advice.